NTIRE 2025 Challenge on Day and Night Raindrop Removal for Dual-Focused Images: Methods and Results –Supplementary Materials–

Xin Li† Yeying Jin† Xin Jin† Zongwei Wu† Bingchen Li† Yufei Wang† Wenhan Yang† Yu Li† Zhibo Chen† Bihan Wen† Robby T. Tan† Radu Timofte† Qiyu Rong Hongyuan Jing Mengmeng Zhang Jinglong Li Xiangyu Lu Yi Ren Yuting Liu Meng Zhang Xiang Chen Qiyuan Guan Jiangxin Dong Jinshan Pan Conglin Gou Qirui Yang Fangpu Zhang Yunlong Lin Sixiang Chen Guoxi Huang Ruirui Lin Yan Zhang Jingyu Yang Huanjing Yue Jiyuan Chen Qiaosi Yi Hongjun Wang Chenxi Xie Shuai Li Yuhui Wu Kaiyi Ma Jiakui Hu Juncheng Li Liwen Pan Guangwei Gao Wenjie Li Zhenyu Jin Heng Guo Zhanyu Ma Yubo Wang Jinghua Wang Wangzhi Xing Anjusree Karnavar Diqi Chen Mohammad Aminul Islam Hao Yang Ruikun Zhang Liyuan Pan Qianhao Luo XinCao Han Zhou Yan Min Wei Dong Jun Chen Taoyi Wu Weijia Dou Yu Wang Shengjie Zhao Yongcheng Huang Xingyu Han Anyan Huang Hongtao Wu Hong Wang Yefeng Zheng Abhijeet Kumar Aman Kumar Marcos V. Conde Paula Garrido Daniel Feijoo Juan C. Benito Guanglu Dong Xin Lin Siyuan Liu Tianheng Zheng Jiayu Zhong Shouyi Wang Xiangtai Li Lanqing Guo Lu Qi Chao Ren Shuaibo Wang Shilong Zhang Wanyu Zhou Yunze Wu Qinzhong Tan Jieyuan Pei Zhuoxuan Li Jiayu Wang Haoyu Bian Haoran Sun Subhajit Paul Ni Tang Junhao Huang Zihan Cheng Hongyun Zhu Yuehan Wu Kaixin Deng Hang Ouyang Tianxin Xiao Fan Yang Zhizun Luo Zeyu Xiao Zhuoyuan Li Pham Hoang Le Nguyen Dinh Thien An Luu Thanh Son Kiet Van Nguyen Ronghua Xu Xianmin Tian Weijian Zhou Jiacheng Zhang Yuqian Chen Yihang Duan Yujie Wu Suresh Raikwar Arsh Garg Kritika Jianhua Zheng Xiaoshan Ma Ruolin Zhao Yongyu Yang Yongsheng Liang Guiming Huang Qiang Li Hongbin Zhang Xiangyu Zheng A.N. Rajagopalan

1. Teams and Methods

1.1. Miracle

This team proposes the STRRNet [39], which is developed based on Restormer [50], as shown in Fig 1. They categorize the training images into four classes based on lighting conditions and raindrop types:

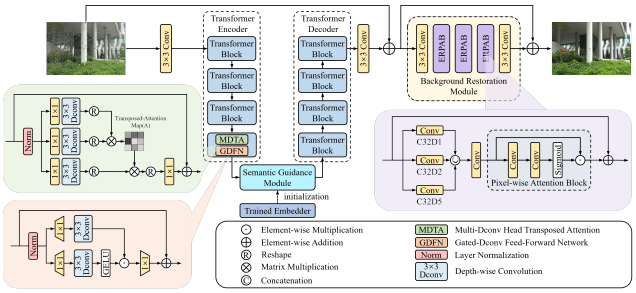

Figure 1. The framework of STRRNet, proposed by Team Miracle

night_bg_focus, night_raindrop_focus, day_bg_focus, and day_raindrop_focus. As shown in Fig 2, a text embedder is first trained on the labeled training set using these

*X. Li, Y. Jin, X. Jin, Z. Wu, B. Li, Y. Wang, W. Yang, Y. Li, Z. Chen, B. Wen, R. Tan and R. Timofte are the challenge organizers

The other authors are participants of the NTIRE 2025 Challenge on Day and Night Raindrop Removal for Dual-Focused Images.

The NTIRE2025 website: https://cvlai.net/ntire/2025/

The Competition website https://codalab.lisn.upsaclay.fr/competitions/21345

The Raindrop Clarity database: https://github.com/jinyeying/RaindropClarity

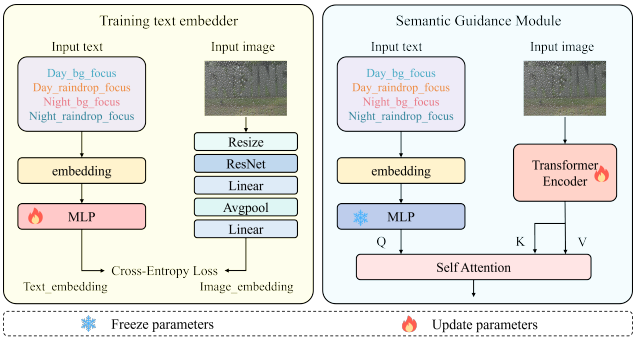

Figure 2. Semantic Guidance Module of STRRNet, proposed by Team Miracle

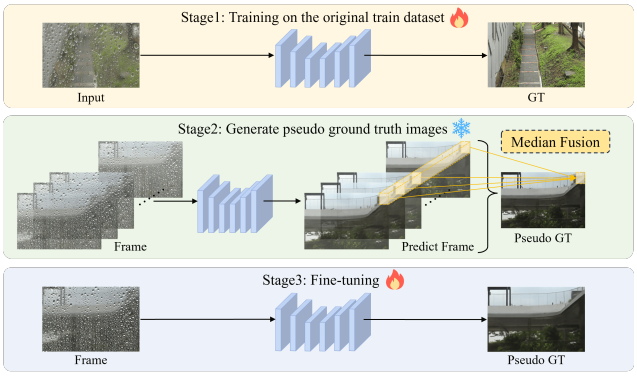

Figure 3. Training Strategy of STRRNet, proposed by Team Miracle

four categories. Then, they design a semantic guidance module, which is added at the end of the Restormer encoder. This module utilizes the encoded image features from Restormer to guide the decoder in performing distinct image restoration operations for the four different types of images. Additionally, they introduce a background restoration subnetwork at the output of Restormer, which consists of multiple convolutional layers to enhance image details. The training strategy of STRRNet is illustrated in the Fig 3. First, they train a pre-trained model on the original training dataset. This model is then used to perform inference on all frames within the same scene in the test set. The inferred images may still contain residual raindrops and artifacts. Since the background remains consistent across different time frames while raindrop positions vary temporally, this motivates them to perform median fusion across multiple frames from the same scene to obtain a median-fused image. Due to the dynamic nature of raindrop artifacts, their inconsistent locations across frames typically prevent them from appearing in the median values, whereas the stable background information is preserved. They then treat the median-fused image as a pseudo ground truth and use it in a semi-supervised fine-tuning phase to enhance the model's raindrop removal capability on unlabeled images.

Training description The Adam optimizer is used for training, with a total of 500,000 iterations. The learning rate is set to 0.0003 for the first 9,2000 iterations and then gradually decays from 0.0003 to 0.000001 for the remaining iterations. Images are randomly cropped to a fixed size of 128×128 for network training, and geometric image augmentation is applied. The network is optimized using the L1 loss function and the multi-scale SSIM loss function, with weights of 1 and 0.2, respectively. All their experiments were conducted on an RTX 4090.

Testing description Use a sliding window to move across the image, applying the model for rain removal on each window. Set the sliding window size to 128×128 with an overlap of 32. Then, obtain a median image by applying median fusion to the images of the same scene. Finally, perform a weighted sum of the median image and the original image to obtain the final output.

1.2. EntroVision

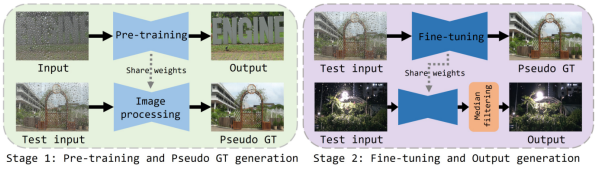

Figure 4. Overview of the technique proposed by Team EntroVision for raindrop removal.

This team utilizes a two-stage approach to achieve raindrop removal in this challenge. The technical overview is shown in the Fig. 4. Given the impact of multi-scale features in image deraining [10, 13], they pre-train the deraining model in the first stage using MSDT [7] on the RainDrop Clarity [20] training set. To enhance the model's generalization capability, additional pre-training is conducted on the UAV-Rain1k [5] dataset. Then, the pre-trained deraining model is utilized to obtain the pseudo-ground-truth images for testing images for the test-time learning of the second stage.

In the second stage of the process, they performed fine-tuning using the test samples and the generated paired pseudo ground-truth images. This approach provides a clear direction for transferring pretrained knowledge, rather than simply relying on the model's generalization ability. Subsequently, they process the testing inputs using this fine-tuned deraining model to obtain the final output results. Notably, they designed the deraining network specifically according to the characteristics of the dataset used. To address scenarios where blurry and clear backgrounds might coexist, they employ median filtering to preserve edge information

and avoid excessive blurring. Through the above two-stage processing strategies and the specially designed median filtering technique, they obtain clear deraining results.

1.3. IIRLab

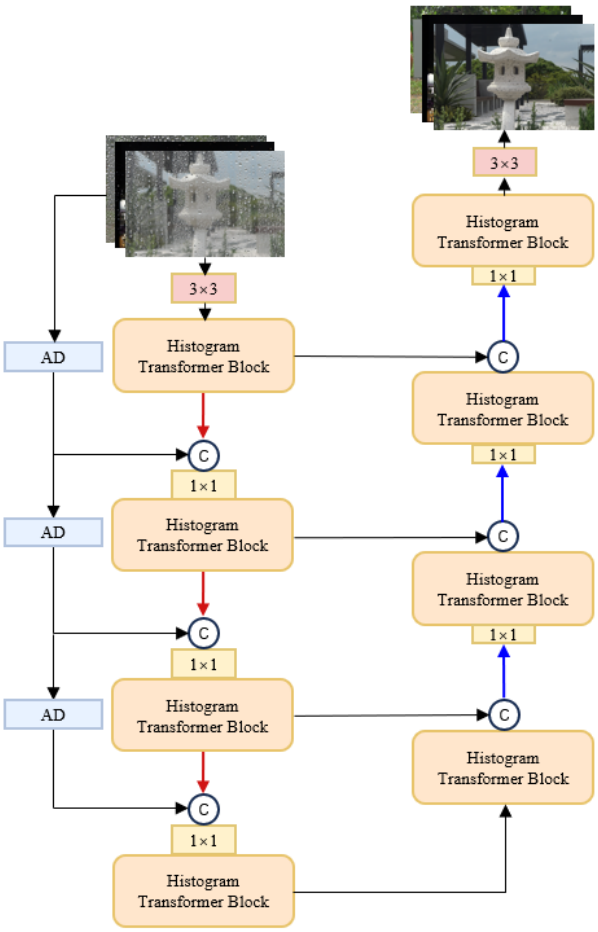

Figure 5. The pipeline of the method proposed by Team IIRLab

This team observed that the raindrop dataset used in this challenge differs significantly from existing raindrop removal datasets. Traditional datasets primarily focus on background clarity during image shooting, resulting in clean backgrounds and blurry raindrops in the foreground. In such cases, simply removing the raindrops is sufficient to recover a clear background. However, the dataset in this challenge includes a notable portion of images where the camera is focused on the raindrops during shooting, leading to clean raindrops and blurred backgrounds. This introduces a new challenge: removing not only the raindrops themselves but also mitigating background blur caused by defocus, to ultimately recover a clean draining image.

Based on this new question, this team has chosen the Histoformer [43] network as their raindrop removal model, as shown in Fig 5. Histoformer is a transformer-based model designed to restore images degraded by severe weather conditions. It incorporates a histogram self-attention mechanism, which sorts and segments spatial features into intensity-based bins and applies self-attention within each bin. This enables the model to focus on spatial features across dynamic intensity ranges and handle long-range dependencies between similarly degraded pixels. Built upon Restormer [50], Histoformer is well-suited for addressing both raindrop artifacts and defocus blur, making it a strong candidate for this task. Based on its architecture and prior performance, the team chose Histoformer as the core model for this challenge.

Training Details. This team utilized all image pairs provided in the training set, consisting of raindrop-degraded images (Drop) and their corresponding clean background images (Clean). From this, they extracted 1,200 image pairs for validation, including 400 daytime and 800 nighttime samples. For training, they employed a two-stage training strategy that combines regular training with subsequent fine-tuning. In the first stage, the draining model was trained for 300,000 iterations using the default Histoformer configuration. In the second stage, the model was fine-tuned for an additional 13,000 iterations using only the $\mathcal{L}_1$ loss function. This progressive training approach effectively enhanced model performance, leading to improved final results.

Implememtation Details. The implementation is based on PyTorch and was conducted on an NVIDIA RTX 3090 GPU. The network was trained for a total of 300,000 iterations, with an initial batch size of 6 and a patch size of 128×128, following a progressive learning strategy. The team employed the AdamW optimizer with an initial learning rate of 3×10^{-4} for the first 92,000 iterations, which was then gradually reduced to 1×10^{-6} using a cosine annealing schedule over the remaining 208,000 iterations. The number of blocks at each stage was set as $L_{i\in1,2,3,4} = 4, 4, 6, 8$, and the channel dimension was fixed at $C = 36$. The channel expansion factor in the DGFF module was set to $r = 2.667$. The number of self-attention heads at each stage was configured as $1, 2, 4, 8$, respectively. For data augmentation, horizontal and vertical flips were applied randomly during training.

1.4. PolyRain

This team initialized and trained a dense X-Restormer model with the given dataset based on Restormer [50]. After the first training stage, they finetuned the model on a larger patch size with different loss functions. The whole framework of this team is shown in Fig. 6.

Training Details. The training process consists of two phases. In the first phase, the patch size of the training image is set to 256, with a batch size of 8, and a total of 3×105

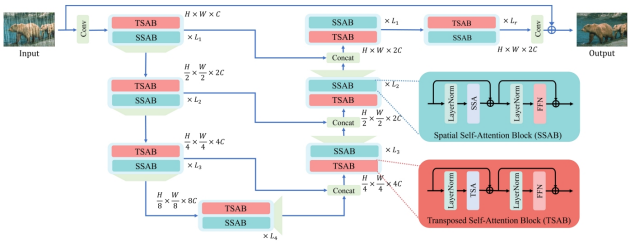

Figure 6. The pipeline of the method proposed by Team PolyRain

training iterations. The learning rate is set to $3e^{-4}$ and $\mathcal{L}_1$ Loss is used as the loss function. In the second phase, the model is fine-tuned with a $5e^{-5}$ learning rate and the image patch with 448 $\times$ 448. During this phase, the model is trained simultaneously using $\mathcal{L}_2$ Loss, LPIPS Loss, and SSIM Loss with weights of 1, 0.1, and 0.1, respectively, for 5$\times$104 iterations.

Testing Details. To enhance the robustness of the model, the self-ensemble technique is employed. The implementation references the BasicSR library .

Implementation Details. This method is implemented based on the famous BasicSR framework [4, 12] written in Python. They utilized the AdamW optimizer with an initial learning rate of $3e^{-4}$. Eight A100 GPUs were used for the model training, lasting for about 72 hours for 300000 iterations. In addition, the CosineAnnealingRestart-CyclicLR scheduler was chosen to restart the learning rate at a setting of $[92000, 208000]$. They did not use any efficient optimization strategy or extra datasets.

1.5. H3FC2Z

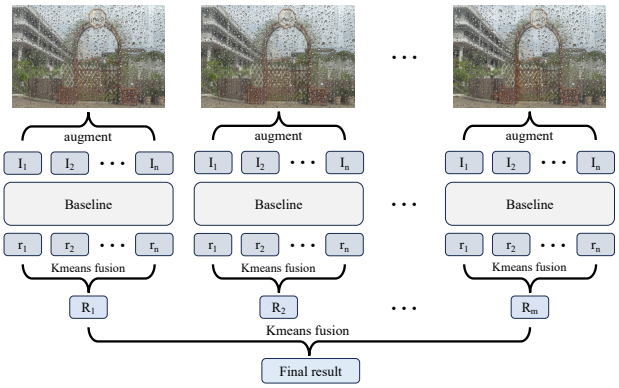

Figure 7. Dual kmeans fusion for RainDrop task proposed by Team H3FC2Z.

This team utilizes the RDDM [29] as the deraining backbone and trained it with a patch size of 256 $\times$ 256. Subsequently, the self-ensemble strategy used in the EDSR [28] was improved by replacing the average ensemble with their "Dual Kmeans fusion" in the Fig. 7. Experimental results indicate that employing "Dual Kmeans fusion" [14] increases the score of the RDDM baseline from 31.72 to 32.56.

Dual Kmeans Fusion. Given a clean image $\mathcal{I}_{gt}$, several raindrop and blur degradations are added to $\mathcal{I}_{gt}$, obtaining m images $(\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m)$. As shown in Fig. 7, taking $\mathcal{I}_1$ as an example, they serve flipping and rotating as augmentations, generating n images $(I_1, I_2, \cdots I_n)$. **Stage-1**: The baseline model processes these images and obtains preliminary results r_i = Baseline(I_i), where $i = 1, 2, \cdots, n$. The images r are flipped or rotated to a fixed angle and Kmeans fusion is performed to obtain R_1. **Stage-2**: After performing the above fusion on images $\mathcal{I}$, they get m results $R_1, R_2, \cdots, R_m$. they perform Kmeans fusion on these m results to get the final result image.

Training and Testing Details. The training dataset provided in this challenge is used for model training. To improve generalization, data augmentation techniques, including rotation and flipping are applied. The model is trained for 100,000 iterations using the AdamW optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.95$, on a single NVIDIA A6000 GPU. Training is conducted with a batch size of 8, a learning rate of 3×10^{-4}, and a patch size of 256$\times$256. During inference, the authors adopt a Dual K-means fusion strategy to further enhance the model's performance. Experimental results indicate that employing "Dual Kmeans fusion" increases the score of RDDM [29] baseline from 31.72 to 32.56. Specifically, Kmeans fusion in stage one increases the score of baseline from 31.72 to 32.26, and Kmeans fusion in stage two increases the score of baseline from 32.26 to 32.56.

1.6. IIC Lab

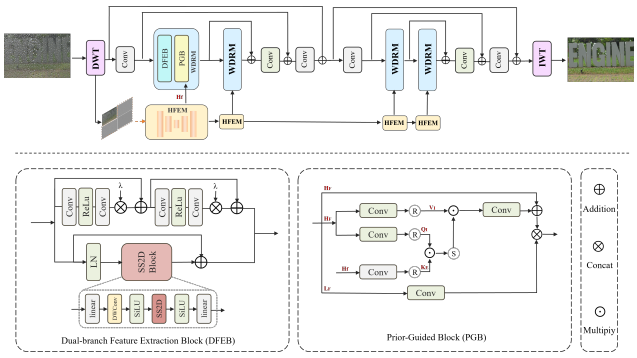

Figure 8. The pipeline of the FA-Mamba proposed by Team IIC Lab.

This team developed an effective frequency-aware and Mamba-based network for image deraining, named FA-Mamba, as shown in Fig. 8. Specifically, the key component

https://github.com/XPixelGroup/BasicSR

of the proposed framework is the Wavelet Domain Restoration Module (WDRM) which contains a Dual-branch Feature Extraction Block (DFEB) that has superior local perception and global modeling capabilities and a Prior-Guided Module (PGM) that provides refined texture detail guidance for feature extraction. It is worth mentioning that the refined texture details are obtained by enhancing the input high-frequency information through the High-Frequency Enhancement Module (HFEM).

Training Details. During training, they utilized the Adam optimizer with a batch size of 1 and a patch size of 256 for a total of 80 iterations. The initial learning rate is fixed at $1e^{-4}$ for 60 iterations, and then decreased to $5e^{-5}$ for 20 iterations. No data augmentation techniques were applied. The entire framework is performed on PyTorch with an NVIDIA GeForce RTX 3090 GPU, which works in an end-to-end learning fashion without costly large-scale pertaining.

1.7. BUPT CAT

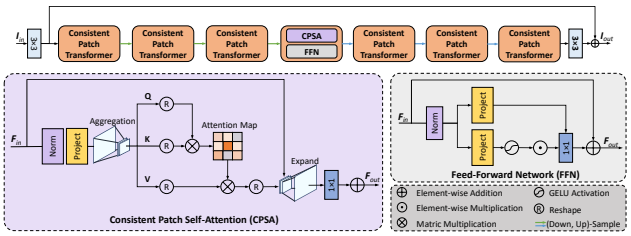

Figure 9. The pipeline of the method proposed by Team BUPT CAT.

As shown in Fig. 9, this team introduces a novel Consistent Patch Transformer (CPT) for dual-focused day and night raindrop removal task, which leverages a UNet-based architecture designed to enhance both spatial consistency and feature representation capability. The framework comprises multiple Consistent Patch Transformer blocks, each consisting of two key components: Consistent Patch Self-Attention (CPSA) and a Feed-Forward Network (FFN). The model utilizes the Test Time Local Converter (TLC) mechanism [15] to effectively revisit global information aggregation and ensure robust and consistent feature learning with different patch sizes at training and testing. The CPSA module is responsible for capturing both long-range dependencies and spatially consistent local details. Instead of using traditional window-based attention mechanisms, the CPSA module integrates a TLC-based feature aggregation and scaling strategy that maintains consistent patch sizes during training and testing, reducing spatial inconsistencies between training and testing.

Training Details. To reduce the training GPU memory, this team augments the input data by randomly cropping the input image into patches of the same size and performing strategies such as random rotation.

Testing Details. During the testing stage, in their self-attention part, they use the TLC strategy to segment and aggregate the full image into a series of patches of the same size as the training patch, and the rest of the model is the full image. This setup can effectively improve the inconsistency of the model's patch size between training and testing, especially in the self-attention part.

Implementation Details. This team utilizes the Pytorch framework with the NVIDIA GeForce RTX 4090. During training, they set the total batch size to 8, the initial learning rate from $5e^{-4}$ to $1e^{-5}$ with a scheduler in 500K iterations, and the patch size is set to 192×192. For the loss function, they use $\mathcal{L}_1$loss and Fourier loss to constrain their model with weights of 1 and 0.1, respectively. They train their framework using the Adam optimizer with $\beta 1 = 0.9, \beta 2 = 0.99$. They set the number of channels to 64 in their network. The total training duration is approximately 80 hours. The training and test sets are official datasets provided by the dual-focused day and night raindrop removal challenge.

1.8. WIRTeam

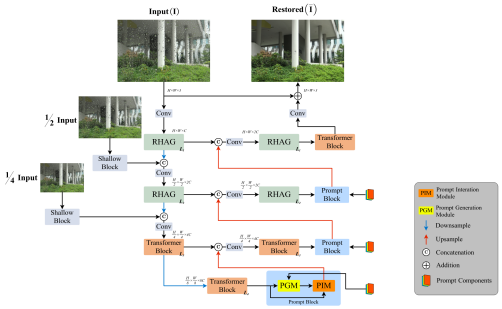

Figure 10. The pipeline of the method proposed by Team WIRteam.

Inspired by recent advancements [11, 26, 37, 50] in image restoration and image deraining, this team propose a novel multi-scale prompt-based image deraining approach (MPID) that incorporates both local and global attention mechanisms, as shown in Fig. 10. Specifically, given a degraded image $I \in R^{H\times W\times 3}$, their model produces a corresponding clear image $\bar{I} \in R^{H\times W\times 3}$ through a 4-level encoder-decoder framework. In the shadow two layers, they employ Residual Hybrid Attention Groups (RHAG) [11] to capture detailed local features. For deeper layers, they integrate Transformer Blocks [50] to facilitate cross-channel global feature extraction. In the encoding phase, acknowledging the advantages of utilizing images at various resolutions for deraining tasks [13], they additionally incorporate multi-scale image information

(at 1/2 and 1/4 of the original resolution) to enhance auxiliary information during the encoding process. During decoding, considering the dual focus on raindrop-focus and background-focus images within the Raindrop Clarity dataset [27]—which introduces both raindrop occlusions and background blurring—they introduce specialized prompt mechanism [26]. These components are designed to address and decouple different types of degradation factors. Notably, Prompt Block, which integrates Prompt Generation Module (PGM) and Prompt Interaction Module (PIM), utilizes image-specific cues to effectively guide the image reconstruction process, thereby improving the clarity and quality of the restored images.

Training Details. They employ the end-to-end training methodology, training their model for 400 epochs. The training procedure is based on the AdamW optimizer with the decay parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is 2e-4 and gradually reduces to 1e-6 with the cosine annealing strategy. Horizontal and vertical flips are adopted for data augmentation. Furthermore, their training is merely based on the Raindrop Clarity dataset provided by the competition organizers, without the use of any extra datasets or pretrained models.

Testing Details. During the testing phase, they preprocess the input images by padding them to the multiple of 32. This ensures that the dimensions of the input images are compatible with the architecture of their method. They don't resort to any other means of Test-Time Augmentation during the testing phase.

1.9. GURain

This team addresses the dual degradation challenge—rain and blur—by unifying the training data into a single framework using the vanilla Restormer, originally designed for single degradation restoration, and by introducing a dynamic rain-aware weighted L1 loss, as shown in Fig. 11. Instead of employing a dual-input network that struggles to differentiate between blur and rain, they merge rainy and blurry images into one input paired with a clear ground truth, allowing Restormer to learn a common mapping for both degradations. Moreover, recognizing that rain streaks affect only parts of an image, their adaptive loss function assigns higher weights to rainy regions and lower weights to clean areas, thereby guiding the network to focus on the more challenging parts of the image. This integrated approach leads to effective deraining while simultaneously mitigating blur, resulting in improved overall image restoration performance as shown in Fig. 12.

Training and Testing Details. They train the default Restormer [50] using a custom dynamic, rain-aware weighted $\mathcal{L}_1$ loss on merged rainy/blurry inputs paired with clear ground truths, with progressive patch size scaling. The custom loss emphasizes challenging rainy regions, yielding improved quantitative metrics and visually cleaner, less blurred results while preserving the effi-

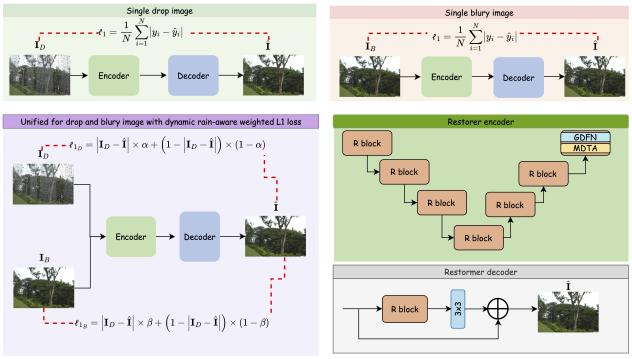

Figure 11. Comparison between the single-degradation Restormer and the dual-degradation method proposed by Team GURain. The top row shows the original Restormer handling drop- or blur-degraded images separately using standard L1 loss. The method by Team GURain (bottom left) unifies both degradations within a single encoder-decoder framework and introduces a dynamic, rain-aware weighted L1 loss to better emphasize challenging regions. The architecture (bottom right) retains the Restormer backbone while enhancing its ability to handle both degradations jointly.

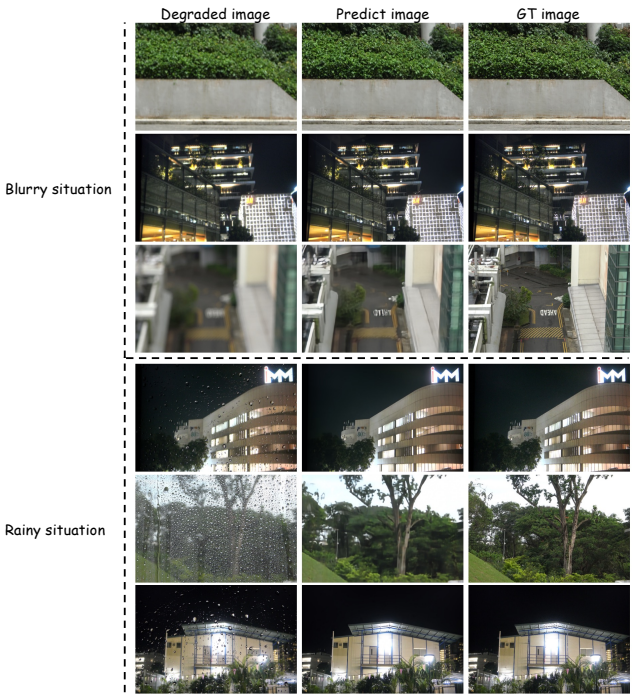

Figure 12. Visualization of the method proposed by GURain on validation image. The top three rows are about deblurry. The bottom three rows are about deraining. Columns from left to right are degraded image, predicted image and ground truth image.

ciency of Restormer [50]. During inference, the standard Restormer pipeline processes single degraded images.

1.10. BIT_ssvgg

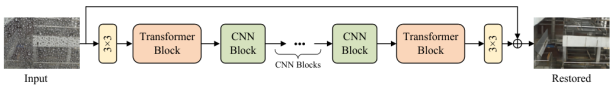

Figure 13. The pipeline of the method proposed by Team BIT_ssvgg.

Figure 14. Visualizations of partial image restoration results by Team BIT_ssvgg.

As illustrated in Fig 13, their method is designed to operate on rain-degraded images and generate the corresponding clean outputs. It follows an encoder–decoder architecture. Given an input degraded image, they first apply a 3×3 convolution to extract initial features, which are then processed by a Transformer module with a global receptive field. Since the rain streaks often appear in large and unevenly distributed regions, it is essential for the network to capture long-range dependencies. To address this, they adopt the attention mechanism from Restormer [50], which enables efficient global context modeling with reduced computational complexity. Subsequently, the features are fed into a series of CNN-based modules to enhance local feature representation and compensate for the Transformer's limited ability to model fine-grained structures. Their CNN blocks are built upon NAFNet [9], chosen for its lightweight design and effectiveness. After that, the features are further refined through another Transformer block and finally passed through a 3×3 convolution layer. The output is then added to the input image in a residual manner to produce the restored image. To preserve spatial information during encoding and decoding, they employ pixel-unshuffle and pixel-shuffle operations for downsampling and upsampling, respectively, after each CNN block. This prevents information loss during resolution changes. By integrating both Transformer and CNN components, their hybrid architecture leverages the strengths of each: global context aggregation and local detail preservation. Moreover, due to its improved generalization ability and reduced tendency to overfit on small datasets, the model can be effectively trained solely on the provided benchmark dataset without requiring any additional data.

Implementation details. Their model is implemented in Python using the PyTorch framework (version 1.13.1). The training is conducted on NVIDIA RTX 4090. They adopt the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The learning rate is scheduled using a cosine annealing strategy, gradually decaying from $1 \times e^{-3}$ to $1 \times e^{-7}$ over the course of training. The training dataset is exclusively provided by the competition organizer, and no additional external data is used. They trained the model for 300 epochs (approximately 30 hours) with a batch size of 8 and a patch size of 256×256. Standard data augmentation techniques are applied, including random horizontal/vertical flipping and random rotations to improve generalization.

1.11. CisdiInfo-MFDehazNet

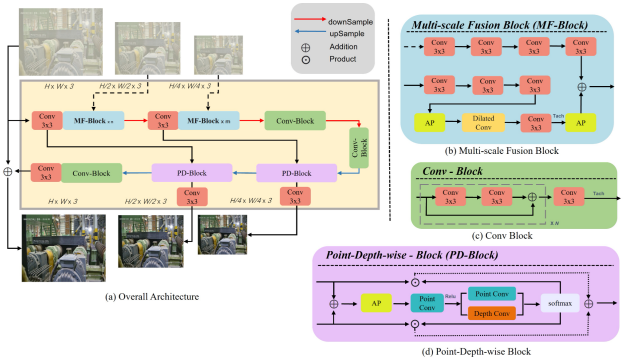

Figure 15. Framework of the method proposed by Team CisdiInfo-MFDehazNet.

In Fig. 15, this team proposes an effective and lightweight model that can be applied in the industrial site, named MFDehaz-Net, which has two specially designed components, *i.e.*, Multi-scale Fusion Block and Point-Depth wise Block, helping it achieve deep fusion of image features of different scales to obtain better global understanding following [18, 41].

Training description The model proposed is implemented with Pytorch1.10.2+CUDA11.3 and trained for 1000 epochs on an NVIDIA GeForce RTX3090 GPU. The batch size and learning rate are set as 16 and 0.0008,respectively, the number of warmup epochs is 50, and Adam is selected as the optimizer. Random rotation and horizontal

inversion are also used as data augmentation methods.

Testing description During testing, all the images are resized to 720x480 and then fed into the loaded pre-trained model. MFDehaz-Net has two specially designed components, i.e., Multi-scale Fusion Block (MF-Block) and Point-Depth wise Block (PD-Block), which help MFDehaz-Net achieve deep fusion of image features of different scales to obtain better global features. Besides, to reduce the damage to the original color of the image caused by image dehazing, MFDehaz-Net integrates the supervision signal in the frequency domain with a specially designed loss function.

1.12. McMaster-CV

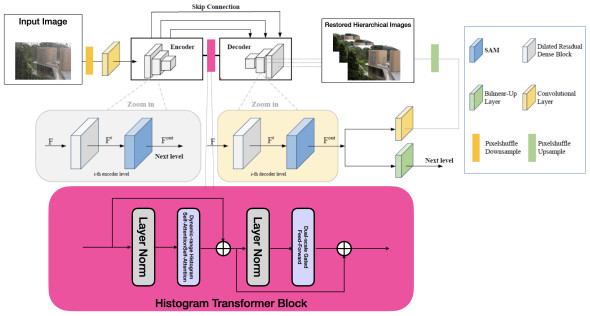

Figure 16. Framework of the method proposed by Team McMaster-CV.

As shown in the Fig. 16, the framework of this team is based on ESDNet [48]. The backbone primarily consists of an encoder-decoder network. At each encoder and decoder level, a Semantic-Aligned Scale-Aware Module (SAM) is incorporated to address scale variations. Additionally, this team introduces a Histogram Transformer Block [42], which employs histogram self-attention with dynamic range spatial attention. This block is placed between the encoder and decoder to achieve global and efficient degradation removal.

Besides, for network training, they designed their loss function based solely on a single level, $\hat{I}_1$, which corresponds to the original image resolution:

$$L_{loss} = L_C\left(I, \hat{I}\right) + \lambda L_{Percep}\left(I, \hat{I}\right) \quad (1)$$

where L_1 and L_{Percep} represent Charbonnier loss [25], and perceptual loss [21], respectively. The weighting factor is set as $\lambda = 0.04$.

Training and Testing Details. This team trained their model on a single NVIDIA 1080Ti(12GB VRAM) GPU. During training, They set the batch size to 3 and the patch size to 480. For the training strategy, following Restormer[50], they adopted the *CosineAnnealingRestartCyclicLR* scheduler, which adjusts the learning rate using a cosine annealing schedule with restarts to promote better convergence and escape local minima. Specifically, the training consists of two cycles with periods of 46,000 and 104,000 iterations, respectively, with minimum learning rates of 0.0003 and 0.000001. The learning rate is reset to its initial value at the beginning of each cycle. To enhance generalization, they incorporated *mixup*-based data augmentation, where training samples are linearly combined using a Beta distribution with a shape parameter of 1.2. Additionally, they enabled the use of identity mapping to retain some original samples during training. The generator is optimized using the *Adam* optimizer with a learning rate of 2×10^{-4} and standard momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). During the testing phase, they directly fed the input into their model to obtain the final output.

1.13. Falconi

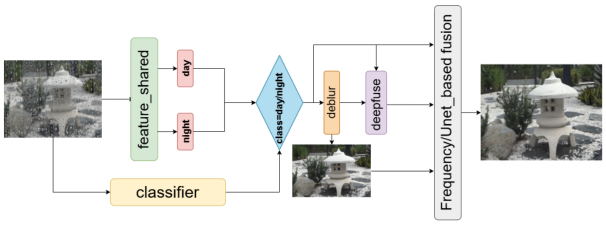

Figure 17. The proposed unified framework for joint image deraining and deblurring by Team Faconi. The framework is composed of four stages: Context Classification, Adaptive Deraining, Multi-Task Feature Extraction, and Iterative Deblurring with Adaptive Fusion.

This team presents a unified framework for image deraining that synergistically combines classification, multi-task learning, and adaptive fusion. Their approach leverages pre-trained models to manage complexity and enhance performance. Specifically, a MobileNetV2-based classifier initially categorizes images as day or night, guiding subsequent processing. A pre-trained Diffusion Transformer(DiT), fine-tuned for both day and night scenarios, serves as a core component for night image adaptation. For deblurring, they incorporate FFTformer. Finally, they explore both traditional and U-Net-based fusion strategies to combine intermediate outputs. The entire framework is trained using a combination of curated external datasets for day/night classification and DiT adaptation, alongside dedicated datasets for deblurring and synthetically generated samples from intermediate deraining stages.

As shown in Fig. 17, their training pipeline comprises four interconnected stages:

Stage 1. Day/Night Classification. They initialize a

MobileNetV2-based classifier to discriminate between day and night scenes, providing a foundational context for subsequent stages.

Stage 2. Adaptive DiT Model. They extend a pre-trained DiT [36] model, initially designed for night imagery, by fine-tuning it on a mixed dataset of day and night images. This stage incorporates the pre-trained classifier (from Stage 1) with a dynamic weighting strategy, enabling context-aware learning based on the time of day.

Stage 3. Multi-Task Branching Network. They construct a three-branch multi-task network. The MobileNetV2 classifier (from Stage 1) forms one branch. The penultimate layer of the adapted DiT model (from Stage 2) serves as a shared feature backbone. Two parallel, final-layer branches, specialized for day and night conditions respectively, are instantiated. During training, the shared backbone and the classification network remain frozen; optimization is restricted to the day- and night-specific branches. The classifier output directly dictates the active branch for a given input.

Stage 4. Iterative Deblurring Refinement. They integrate the FFTformer deblurring network [24] using a two-round fine-tuning process. The first round utilizes a dataset of blurred images. The second round leverages a synthetic dataset generated from the deraining outputs of the DiT model (from Stage 2), thereby aligning the deblurring process with the specific characteristics of the derained images.

During inference, the input image undergoes a two-stage process: deraining followed by deblurring. First, the image is processed by the deraining pipeline (Stage 3), and subsequently enhanced by the deblurring network (Stage 4). To effectively integrate the complementary information from these stages, they employ a dual-fusion approach:

Frequency-Domain Fusion. This method combines derained and deblurred images by adaptively weighting their frequency components based on sharpness, noise, and edge information, prioritizing low-frequency content from the derained image and high-frequency details from the deblurred image.

Learned Fusion. This method utilizes a U-Net-based neural network [30], trained on a diverse dataset of model outputs, to learn a non-linear mapping that optimally fuses the derained and deblurred images, implicitly addressing the weaknesses of each individual processing stage. The table 1 presents the final evaluation results of their image deraining and deblurring framework, with validation and test values reported. The validation score of 33.05 indicates the model's strong generalization capability during the training phase, while the test score of 31.71 reflects its performance on previously unseen data. The relatively small difference between the validation and test results suggests that the model has effectively avoided overfitting, main-taining robust performance across different datasets.

Dataset	Value
Validation	33.05291
Test	31.70666

Table 1. Final evaluation results of the Image Deraining and Deblurring Framework by Team Falconi.

These results validate the effectiveness of the unified framework, which integrates day/night classification, adaptive DiT model, multi-task branching, and iterative deblurring refinement.

This team developed their method based on the Dit model, utilizing the Adam optimizer with a learning rate of $1 \times e^{-5}$, and executed the training process on a single NVIDIA 4090D GPU (24GB). The proposed Raindrop Clarity dataset was exclusively employed, with pairs of (drop, clear) used in stages 1 and 2, and pairs of (blur, clear) utilized in stage 3. In Stage 1, they trained a binary classifier, achieving convergence in under 30 minutes. Stage 2 required 25 hours of training, while Stage 3 took 14 hours. Finally, Stage 4 involved training a U-Net, which was completed in less than 10 minutes. Throughout the training process, learning rate optimization was employed as part of the training strategy.

1.14. Dfusion

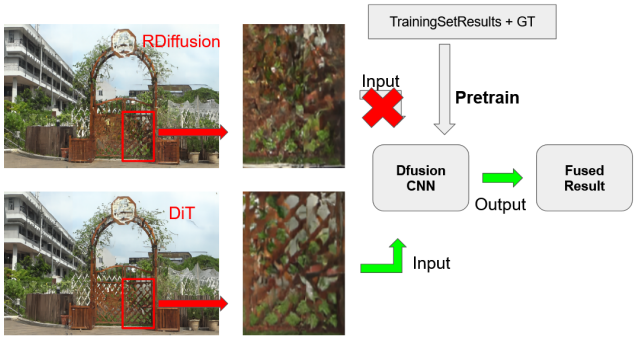

Figure 18. The implementation structure of Team Dfusion.

The proposed **Dfusion** method is a dual-branch fusion framework that leverages the complementary strengths of two state-of-the-art pretrained models—RDiffusion [32] and DiT [36]—to effectively remove raindrops from dual-focused images. As illustrated in Fig. 18, the method first generates two intermediate restored images, each excelling in different visual aspects (for example, one may deliver smooth textures while the other preserves fine details).

These intermediate outputs are then stacked and passed through a lightweight fusion CNN, which is specifically trained to combine the best qualities of both inputs. The final model is optimized using a joint loss function that in-

tegrates PSNR, SSIM, and LPIPS, ensuring both high quantitative accuracy and strong perceptual quality.

To enhance computational efficiency, input images are processed in overlapping patches. These patches are restored separately by the pretrained models and then reassembled prior to the fusion step. This patch-based strategy not only speeds up inference but also helps preserve local details across the image.

Their fusion approach draws inspiration from mixture-of-experts strategies [17], adapting them to the raindrop removal task. By fusing complementary features from two high-performance models, Dfusion is able to achieve robust restoration results in both daytime and nighttime scenarios, as shown in Fig. 19.

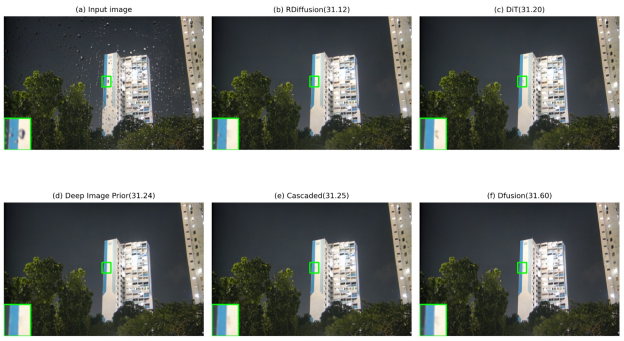

Figure 19. The result comparison of Team Dfusion.

Training Details. The fusion network is trained using a single RTX4060 8GB laptop. Adam optimizer is used with an initial learning rate of 1×10^{-4}. During training, each input image (of resolution 720×480) is split into six overlapping patches of size 256×256. The model is then optimized with a joint loss function that combines PSNR, SSIM, and LPIPS.

Testing Details. In the inference stage, intermediate outputs from RDiffusion and DiT are first generated. These outputs are concatenated into a multi-channel tensor and subsequently processed by the lightweight fusion CNN. This process results in a final restored image that effectively preserves both global structure and fine local details.

1.15. RainMamba

The team proposes a video restoration framework [47] to adapt state space models to Day and Night Raindrop Removal tasks, as shown in Fig. 20. Their approach applies a global scanning mechanism to causally process the temporal data with linear complexity. The core innovation of their method lies in equipping State Space Models (SSMs) with a Hilbert scanning mechanism, which achieves localized scanning across both temporal and spatial dimensions. Given a sequence of rainy video frames, their cascading Coarse-to-Fine Mamba Module (CFM) receives the encoded features as input and causally models temporal corrections using improved state space models. The CFM employs Global Mamba Block (GMB) and Local Mamba Block (LMB) to capture sequence-level global and local spatio-temporal dependencies. They develop a novel Hilbert scanning paradigm in LMB to promote the Mamba's locality learning. To enhance the visual quality of the restored results, they adopt a combination of PSNR loss and perceptual loss in their training process.

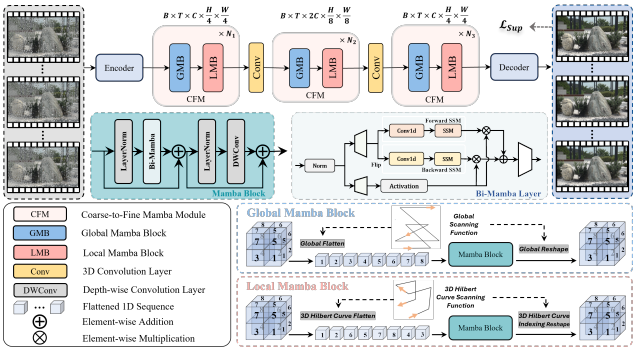

Figure 20. Network architecture of Team RainMamba.

Training Details. Their network is trained on NVIDIA RTX 4090 GPUs and implemented on the Pytorch platform. At each training iteration, the input frame is randomly cropped to a spatial resolution of 256×256, and the number of frames per video clip is 2. The total number of training iterations is 300k. They adopt the Adam optimizer and the polynomial scheduler with a power of 1.0. The initial learning rate of their network is set to 5×10^{-4} with a batch size of 8 and a warm-up start of 2k iterations.

Testing Details. In the testing phase, they take two frames of each video as a segment as input and input the full size of each frame.

1.16. RainDropX

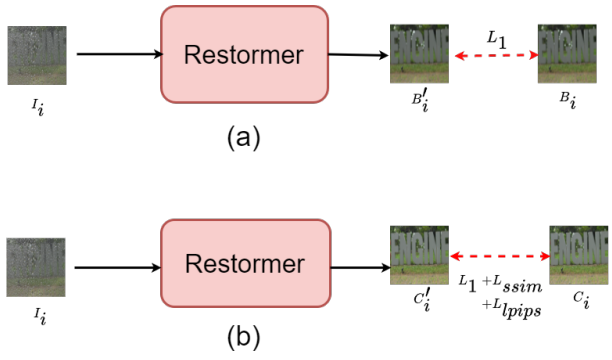

Figure 21. The diagram of the proposed method by Team RainDropX.

The team proposes an enhanced version of Restormer [50], introducing additional constraints for improving the perceptual and structural integrity of the restored images. Unlike the original Restormer model, which was trained for individual restoration tasks using L_1 loss, the proposed method incorporates LPIPS [51] and SSIM [45] losses to preserve perceptual quality, edge preservation, and texture consistency. The training process is conducted in two stages, starting with fine-tuning the pretrained weights for blurred output prediction and then transitioning to predicting clean images with added perceptual losses. The overall framework is shown in Fig. 21.

Training Details. The model is trained using the pretrained Restormer weights with a two-stage training approach. In Stage 1, L_1 loss is used to predict blurred outputs from degraded inputs containing both raindrops and blur. In Stage 2, LPIPS and SSIM losses are introduced alongside L_1 loss to optimize clean image restoration. The training is conducted on the Raindrop Clarity dataset [20], and the model is trained using a single GPU.

1.17. Cidaut AI

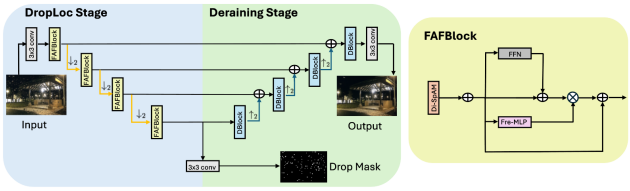

Figure 22. The proposed DropFIR by Team Cidaut AI.

This team proposes DropFIR, an encoder-decoder architecture inspired by DarkIR [16], tailored for raindrop removal, as shown in Fig. 22. The model introduces a DropLoc stage based on a custom Fourier-Attention-Fusion Block (FAFBlock), adapted from FLOL [2], to extract a spatial drop mask. The predicted mask guides the Deraining Stage, enabling the decoder to remove localized drops and blur for image restoration.

Training Details. The model is trained using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01) with a cosine annealing learning rate schedule starting at 10^{-3} and decaying to 10^{-7}. Training is conducted in two phases: a 100-epoch pretraining on 320×480 random crops, followed by 300 epochs on full-resolution images. The model is trained on four NVIDIA H100 GPUs with a batch size of 16, using only the Raindrop Clarity dataset [20].

1.18. DGL_DeRainDrop

This team proposes a Global Semantic Attention (GSA) based two-step dual-focused image raindrop removal method (GSA2Step), as depicted in Fig. 23. They adopt a two-step approach where they break down the dual-focused

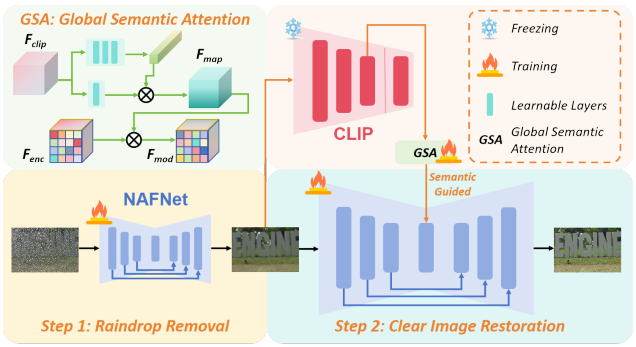

Figure 23. Architecture of the proposed GSA2Step framework by Team DGL_DeRainDrop.

image raindrop removal task into two sub-tasks: raindrop removal and clean image restoration. In the first step, they use a 32-width NAFNet [9] to remove focused and defocused raindrops. In the second step, they employ a Global Semantic Attention (GSA) module that utilizes CLIP [38] semantic features to guide a 64-width NAFNet in reconstructing the final clean image from the background obtained in step 1. The GSA mechanism extracts semantic features using the CLIP image encoder and effectively integrates them into the encoded features obtained from the NAFNet encoder to guide the subsequent decoding process. This architecture addresses various degradations including focused and defocused raindrops, as well as defocused backgrounds.

Training Details. They employ a multi-stage training approach to reduce the network's learning difficulty. Initially, they pretrain the 32-width NAFNet using the Drop and Blur subsets from the training dataset with only L1 loss. Subsequently, they freeze the pretrained NAFNet and train the entire framework using the Drop and Clear subsets, still with only L1 loss. Finally, they unfreeze the parameters of the first NAFNet and jointly fine-tune the entire framework using a combination of L1 loss, SSIM loss, and perceptual loss [21] with weight coefficients of 1.0, 0.5, and 0.01, respectively. The model is trained using the Adam [22] optimizer with learning rate initially set to $4 \times e^{-4}$ and halved after every 200,000 iterations, with the final fine-tuning using a fixed learning rate of $1 \times e^{-5}$. Training is conducted on a single NVIDIA RTX 4090 GPU with PyTorch [33] for approximately 4 days over 1000 epochs with no additional datasets utilized.

1.19. xdu_720

This team adopts a two-stage architecture combining a mask prediction network and a Transformer-based restoration model, as depicted in Fig. 24. In the first stage, a residual convolutional network is used to predict the raindrop mask from the input raindrop image, supervised by an

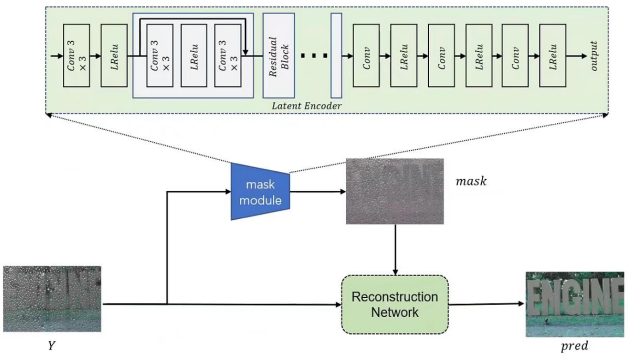

Figure 24. The overall framework by Team xdu_720.

L1 loss using ground truth masks obtained from the difference between the Raindrop and Blur images. In the second stage, the predicted mask is fused with features within a Transformer-based deraining network (FFTformer [23]), where mask features are concatenated after each block. The final output is optimized with an MSE loss against the Clear image.

Training Details. Training is conducted in two stages. Stage one trains the mask prediction network for 20 epochs using L1 loss. Stage two trains the mask-prior-guided Transformer restoration network for 80 epochs using MSE loss. The predicted mask from stage one is used as additional guidance throughout the Transformer layers. No additional datasets or pre-training are used.

Testing Details. A subset of 3,000 image pairs from the training set is used for testing. During inference, the raindrop image is processed jointly by the mask prediction network and the FFTformer [23]. The evaluation is based on PSNR, SSIM, and LPIPS metrics.

1.20. EdgeClear-DNSST

This team proposes EdgeClear-DNSST, a transformer-based encoder-decoder model designed for raindrop removal in dual-focused images. The model integrates Sparse-Sampling Attention for efficient long-range dependency modeling, combined with specialized modules including RaindropEdgeEnhancer, PyramidAttention, and RaindropFeatureModulation to address fine-grained raindrop degradation. The network adopts a multi-stage structure [50] with four encoder and decoder stages, enhanced by skip connections and multi-scale feature fusion.

Training Details. The model is trained from scratch using only the day and night images from the RaindropClarity dataset [20], split into a 9:1 training-validation ratio. The training process utilizes the Adam optimizer with an initial learning rate of 0.0006, scheduled by cosine annealing (T_{max}=100, η_{min}=1 $\times$ e^{-6}). The batch size is 24 per GPU with gradient accumulation (accumulate grad batches=2). Data augmentation includes random horizontal flips and 90° rotations. A combination of Y-channel PSNR loss, SSIM loss, LPIPS perceptual loss, and difference-weighted loss is used. An early stopping strategy with a patience of 20 is applied based on validation PSNR.

Testing Details. During inference, they employ a sliding window strategy with a window size of 128×128 and a 32-pixel overlap to handle images of arbitrary resolution. A weighted averaging method is used at the window edges to suppress boundary artifacts. Bilinear interpolation with aligned corners is adopted to mitigate edge color anomalies.

1.21. MPLNet

This team builds upon the multi-stage fusion network MPRNet [49] and enhances its performance under non-uniform illumination conditions by integrating a Global Illumination Detection Module (IllumAdjust) [3, 46]. IllumAdjust generates a single-channel illumination map through a series of convolution and attention layers, which is then embedded into the encoder-decoder framework via a custom-designed UniMetaFormer unit. The UniMetaFormer module, based on Metaformer [53], replaces the CAB modules in MPRNet's encoder and decoder. It comprises a Dynamic Tanh normalization layer (DyT) [55], an Illumination-guided Channel Attention Block (ICAB), and a convolution-based MLP. ICAB integrates channel and spatial attention mechanisms guided by the illumination map, enhancing the model's adaptability to illumination variations. The network retains MPRNet's three-stage progressive restoration and employs Supervised Attention Modules (SAM) for stage-wise feature fusion. An Original Resolution Block (ORB) is used in the final stage to further refine the output.

Training Details. This team trained their model using the same dataset as in [49]. The loss function combines Charbonnier Loss, Edge Loss, and LPIPS Loss to balance pixel-level accuracy and perceptual quality. They performed validation after each epoch and saved model checkpoints based on the best PSNR and LPIPS scores, as well as at fixed intervals. Gradient clipping was applied to stabilize training. All experiments were conducted on an NVIDIA GPU.

1.22. Singularity

Figure 25. The overall pipeline proposed by Team Singularity.

This team proposes a two-step pipeline to address raindrop and blur removal challenges, as depicted in Fig. 25. The approach leverages the distinct characteristics of raindrops and blurring artifacts, which require different han-

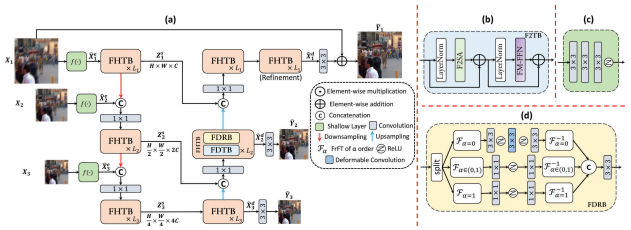

Figure 26. The overall architecture for deblurring proposed by Team Singularity.

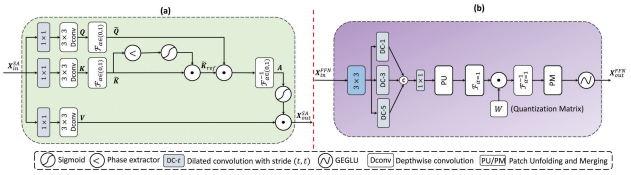

Figure 27. The (a) self-attention mechanism Fractional Frequency aware Self-Attention (F2SA), and the (b) Frequency quantized feed forward network FQ-FFN proposed by Team Singularity.

dling strategies. By separating the tasks into two stages, they apply specialized techniques to effectively remove each type of artifact, leading to higher-quality image reconstruction. For raindrop removal, they utilize FFTformer [23] to effectively handle the irregular patterns and occlusions caused by raindrops. For blur removal, they develop an enhanced version of F2former [35], leveraging Fractional Fourier Transform (FRFT) for non-uniform deblurring, as shown in Fig. 26. The blur removal framework consists of two components: FHTB and FDTB. Specifically, FHTB includes FDRB (similar to F3RB in [35]) and FDTB, which use deformable convolutions to adapt to varying motion or blur patterns, improving edge restoration. FDTB uses an FRFT-based attention mechanism, inspired by [35], and integrates quantization from FFTformer's FSAS module to focus on relevant frequency components, ensuring efficient handling of the blur pattern at multiple levels. The details are illustrated in Fig. 27.

Training Details. This team trains the raindrop and blur removal models separately. The two models are trained following the strategy of NAFNet [8] and F2former [34]. They use the standard data augmentation, the Adam optimizer with default settings, and a learning rate of 1e-3. They update their model with a cosine annealing strategy for 600,000 iterations. Both models are trained on 256×256 images with a batch size of 8, and a patch size of 8×8 is used for self-attention. They apply L1 loss in both spatial and frequency domains, with adversarial training applied for 50,000 iterations to improve perceptual quality. All experiments are conducted on an 80GB Nvidia A100 GPU, with training configurations modified from NAFNet.

1.23. VIPLAB

This team utilizes an efficient unified framework with a two-stage training strategy to explore the weather-general and weather-specific features separation. The first training stage aims to learn the weather-general features by taking the images under various weather conditions as inputs and generating the coarsely restored results. The second training stage aims to learn to adaptively expand the specific parameters for each weather type in the deep model, where the requisite positions for expanding weather-specific parameters are automatically learned. Finally, they adopt NAFNet [9] to enhance the textures.

Training Details. They optimize the model for 200 epochs using AdamW optimizer with a learning rate of 1e-4. They employ data augmentation techniques, including random cropping and flipping.

1.24. 2077Agent

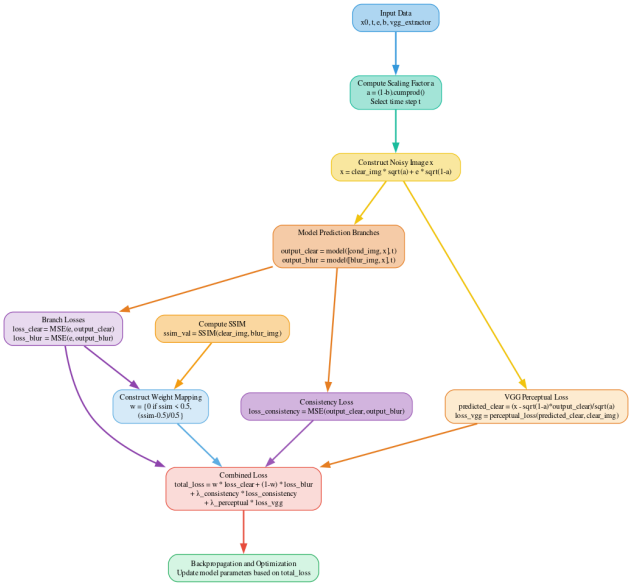

Figure 28. Architecture of the Loss Computation and Optimization Module by Team 2077Agent.

This team builds the model based on the pre-trained DiT [36] model. They propose a cost-effective and efficient fine-tuning approach for dual-focused day-and-night raindrop removal by optimizing the loss function. Specifically, they introduce an adaptive loss function that integrates Structural Similarity (SSIM) and VGG-based perceptual losses. The SSIM-based weighting dynamically emphasizes regions with significant texture differences between clear and blurry images. Meanwhile, the perceptual loss leveraging VGG features ensures visually realistic outcomes. Additionally, a consistency constraint is incorporated to stabilize noise estimation between clear and

blurred branches. This process is illustrated in Fig. 28. They utilize additional datasets, including LHP-Rain [19] and DIV2K [1], for data augmentation to enhance the model's generalization capability to unseen scenarios.

Training Details. They fine-tune the pre-trained DiT [36] model provided by the official repository using the Adam optimizer with an initial learning rate of $2 \times e^{-5}$ for about 50 epochs. They resize the original training images to two scales (720×480 and 360×240) to enhance the generalization ability of the model towards diverse resolutions. They randomly crop the training patches into 64×64 for training. They optimize the DiT model with a linear beta scheduling strategy (ranging from 0.0001 to 0.02) with 1000 diffusion steps. Additionally, they develop a hybrid loss function with SSIM-based adaptive weighting, VGG perceptual loss, and consistency constraints, significantly improving the model's detail recovery capability and visual quality of the restored images.

Testing Details. In the testing phase, they employed an implicit sampling method with 25 sampling steps to achieve efficient and high-quality raindrop removal. All output images are 720×480 pixels. To mitigate boundary artifacts introduced during patch stitching, they utilize a grid-based overlapping strategy with a 16-pixel overlap during image reconstruction. Unlike the training phase, they do not crop patches during testing to ensure a comprehensive global evaluation of the full-resolution images. They conduct all evaluations with a single NVIDIA 4090 GPU.

1.25. X-L

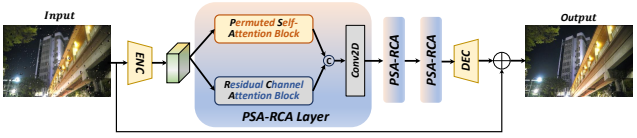

Figure 29. The overview of the proposed method by Team X-L.

This team develops the model following the popular encoder-decoder paradigm. They propose to enhance the middle features by a combination of Permuted Self-Attention blocks (PSA) [54] and Residual Channel Attention blocks (RCA) [52], as shown in Fig. 29. Each PSA-RCA layer comprises two parallel branches: one branch consists of residual channel attention blocks [52], which are designed to capture local features, while the other branch includes Permuted Self-Attention Blocks (PSAB) [54] that model global relationships with low computational complexity. This dual-branch architecture effectively integrates both local and global information, thereby enhancing the precision and efficiency of feature extraction. Following the processing through these three PSA-RCA layers, the refined features are passed to a feature decoder for additional processing and reconstruction, ultimately yielding a clear and detailed output.

Training Details. The model was trained using the provided dataset, with the L1 loss function employed to optimize the training process. No additional datasets were utilized during training. The training was conducted on a single NVIDIA 4090 GPU.

Testing Details This team performed ensemble on the test data, adopting the same strategy as EDSR, which involves rotating the images at different angles to achieve this. This method enhances the robustness and quality of the final output by generating predictions from multiple perspectives. By integrating these predictions from various angles, the model can better capture the details and structural information in the images, significantly improving the performance of the image processing task.

1.26. UIT-SHANKS

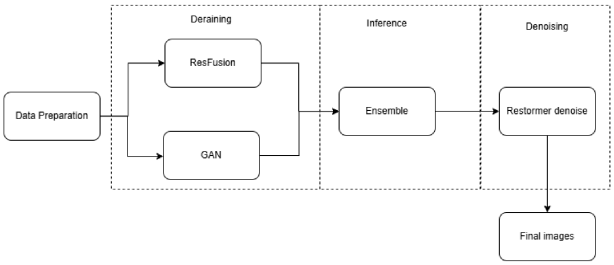

Figure 30. The proposed method of Team UIT-SHANKS.

This team proposes a two-stage approach to handle the challenge. In the first stage, they develop their model based on ReFusion [31]. In the second stage, they further optimize their model by incorporating the GAN-based training strategy, with a semi-supervised dataset derived from test data. They generate the dataset after training the model after 100 epochs. They build the discriminator in the GAN training phase based on UNet [40] architecture. The overall framework is depicted in Fig. 30.

Training Details. In the first stage, they adopt the AdamW optimizer with the CosineAnnealing scheduler to optimize the model. They use MSE loss to obtain a fidelity-oriented model. In the second stage, they adopt the GAN-based training strategy with the standard GAN loss and VGG perceptual loss. Additionally, they leverage multiscale loss, SSIM loss and MASK loss to optimize the model. They finish their training with two T4 GPUs and a P100 GPU from the Kaggle platform.

1.27. One Go Go

This team presents a two-stage approach for day and night raindrop removal for dual-focused images. Their method processes raindrop removal and defocus-blurring

separately. They employ a retrained Restormer [50] model specifically for raindrop removal in the first stage, while utilizing a fine-tuned IPT [6] model to address the defocus-blurring issues present in the second stage. They maintain the same hyperparameters as in the original Restormer paper, with the number of channels set to [48, 96, 192, 384], respectively. For the implementation of IPT, they only retrain its head and tail components while preserving the core transformer blocks fixed.

Training Details. They apply the same training process for both day and night images. However, they split the dataset into two functional subsets: one containing raindrops and one for blur correction. They develop two separate models rather than pursuing an end-to-end solution. The Restormer [50] model was trained from scratch specifically for raindrop removal using the challenge dataset, while the IPT [6] model leveraged pretrained weights with fine-tuning limited to only the head and tail components on the defocus-blurring subset. They optimize the model following the original Restormer approach, while adopting a patch size of 48×48 for IPT training. They accomplish their training with 4 NVIDIA GeForce RTX 3090 GPUs.

Testing Details. Their testing pipeline consists of two sequential processing stages. First, they apply the trained Restormer [50] model to remove raindrops from the input images. Subsequently, they feed these processed images to the fine-tuned IPT [6] model to restore the defocus-blurring. The inference costs approximately 455ms on a single GPU with their proposed testing pipeline.

1.28. DualBranchDerainNet

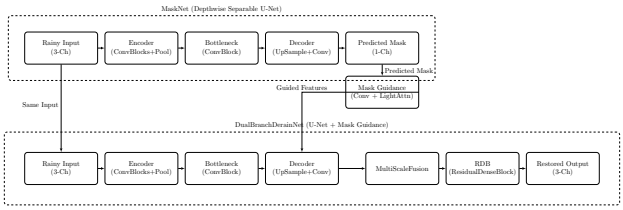

Figure 31. The network architecture proposed by Team DualBranchDerainNet. Zoom in for a better view.

This team proposes a two-stage framework comprising a mask estimation network (MaskNet) and a dual-branch deraining network (DualBranchDerainNet), as shown in Fig. 31. The MaskNet predicts rain masks by learning from the difference between the "rain-focused" and "blur-focused" inputs. The DualBranchDerainNet takes both the input image and the predicted mask, using a lightweight U-Net with multi-scale fusion and channel attention techniques to remove raindrops and restore clear details. They also adopt a WGAN-GP-based discriminator to improve the perceptual quality of the restored results, along with frequency and gradient losses for more faithful restoration.

Training Details. They train the proposed model for 300 epochs using Adam optimizer, with a batch size of 16 and an initial learning rate of 4e-3 on an A100 PCIE 40GB GPU. They adopt a combination of losses to optimize their model, including L1, SSIM, WGAN-GP, frequency domain, and gradient-based losses. The approximate training time to obtain their model is 60 hours.

Testing Details. They evaluate their model by directly inferring the full-resolution images. The final output is a blend of the original input and a learned residual, guided by the predicted raindrop mask.

1.29. QWE

This team develops their method upon the previous backbone, TransWeather [44], by leveraging prior knowledge from the pre-trained model and applying continuous learning on the competition dataset. The model architecture, algorithms, and module structure remain consistent with those outlined in the original TransWeather paper. However, they preprocess the dataset to align with the specific scenarios of this competition and introduce modifications to the loss function and training process.

Training Details. They follow the officially released code of TransWeather [44] and fine-tune the model for 200 epochs with RTX3050 GPUs.

Acknowledgments

This work was partially supported by NSFC under Grant 623B2098 and the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH. We thank the challenge sponsor: Eastern Institute for Advanced Study, Ningbo. This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2025 sponsors: ByteDance, Meituan, Kuaishou, and University of Wurzburg (Computer Vision Lab).

Organizers

Title: NTIRE 2025 Challenge on Day and Night Raindrop Removal for Dual-Focused Images

Members:

Xin Li[1] (xin.li@ustc.edu.cn),
Yeying Jin[2,3] (jinyeying@u.nus.edu),
Xin Jin[4] (jinxin@eitech.edu.cn),
Zongwei Wu[5] (zongwei.wu@uni-wuerzburg.de),
Bingchen Li[1] (lbc31415926@mail.ustc.edu.cn),
Yufei Wang[6] (ywang25@snap.com),
Wenhan Yang[7] (ywang25@snap.com),
Yu Li[8] (liyu@idea.edu.cn),
Zhibo Chen[1] (chenzhibo@ustc.edu.cn),
Bihan Wen[9] (bihan.wen@ntu.edu.sg),

Robby T. Tan[2] (robby.tan@nus.edu.sg),
Radu Timofte[5] (Radu.Timofte@uni-wuerzburg.de)
Affiliations:
[1] University of Science and Technology of China
[2] National University of Singapore
[3] Tencent
[4] Eastern Institute of Technology, Ningbo
[5] Computer Vision Lab, University of Würzburg
[6] Snap Research
[7] Pengcheng Laboratory
[8] IDEA
[9] Nanyang Technological University

Miracle

Title: Semantics-Guided Two-Stage Raindrop Removal Network
Members: Qiyu Rong (20231083510916@buu.edu.cn), Hongyuan Jing, Mengmeng Zhang, Jinglong Li, Xiangyu Lu, Yi Ren, Yuting Liu and Meng Zhang
Affiliations:
Beijing Union University

EntroVision

Title: Two-stage Multi-scale Transformer for Day and Night Raindrop Removal
Members: Xiang Chen[1] (chenxiang@njust.edu.cn), Qiyuan Guan[2], Jiangxin Dong[1], Jinshan Pan[1]
Affiliations:
[1]Nanjing University of Science and Technology
[2]Dalian Polytechnic University

IIRLab

Title: A raindrop removal method based on Histoformer
Members: Conglin Gou[1] (gou_conglin@tju.edu.cn), Qirui Yang[1], Fangpu Zhang[1], Yunlong Lin[2], Sixiang Chen[3], Guoxi Huang[4], Ruirui Lin[4], Yan Zhang[5], Jingyu Yang[1], Huanjing Yue[1]
Affiliations:
[1]TianJin University
[2]Xiamen University
[3]The Hong Kong University of Science and Technology, Guangzhou
[4]University of Bristol, UK
[5]National University of Singapore, Singapore

PolyRain

Title: Finetuning Dense X-Restormer for Image Deraining
Members: Jiyuan Chen[1] (jiyuan.chen@connect.polyu.hk), Qiaosi Yi[1] (qiaosi.yi@connect.polyu.hk), Hongjun Wang[2], Chenxi Xie[1], Shuai Li[1], Yuhui Wu[1]
Affiliations:
[1]The Hong Kong Polytechnic University
[2]The University of Tokyo

H3FC2Z

Title: Dual kmeans fusion for the RainDrop task
Members: Kaiyi Ma, Jiakui Hu (jiakuihu29@gmail.com)
Affiliations:
Xidian University

IIC Lab

Title: FA-Mamba
Members: Juncheng Li[1] (junchengli@shu.edu.cn), Liwen Pan[1], Guangwei Gao[2]
Affiliations:
[1]Shanghai University
[2]Nanjing University of Posts and Telecommunications

BUPT CAT

Title: Consistent Patch Transformer for Dual-Focused Day and Night Raindrop Removal
Members: Wenjie Li (lewj2408@gmail.com), ZhenyuJin, Heng Guo, Zhanyu Ma
Affiliations:
Beijing University of Posts and Telecommunications

WIRTeam

Title: MPID: Image Deraining with Multi-Scale Prompt-Based Learning
Members: Yubo Wang (wangyubo@stu.hit.edu.cn), Jinghua Wang
Affiliations:
Harbin Institute of Technology (Shenzhen)

GURain

Title: Unified Image Restoration for Rain and Blur Using Restormer with a Dynamic Rain-Aware Weighted $\mathcal{L}_1$ Loss
Members: Wangzhi Xing (w.xing@griffith.edu.au), Anjusree Karnavar, Diqi Chen, Mohammad Aminul Islam
Affiliations:
Griffith University

BIT_ssvgg

Title: Hybrid Network of CNN and Transformer for Raindrop removal
Members: Hao Yang (3120235187@bit.edu.cn), Ruikun Zhang, Liyuan Pan
Affiliations:
Beijing Institute of Technology

CisdiInfo-MFDehazNet

Title: MFDehazNet-An Easily Deployable Image Dehazing Model for Industrial Sites
Members: Qianhao Luo (QianHao.Luo@cisdi.com.cn), XinCao (Xin.A.Cao@cisdi.com.cn)
Affiliations: CISDI Information Technology CO., LTD

McMaster-CV

Title: **RainHistoNet**: Single-Image Day and Night Raindrop Removal via Histogram-Guided Restoration
Members: Han Zhou (zhouh115@mcmaster.ca), Yan Min, Wei Dong, Jun Chen
Affiliations:
Department of Electrical and Computer Engineering, McMaster University

Falconi

Title: No Title
Members: Taoyi Wu (taoyiwu81@gmail.com), Weijia Dou, Yu Wang, Shengjie Zhao
Affiliations:
Tongji University

Dfusion

Title: Dffusion: A new method to fuse existing solutions with simple CNN
Members: Yongcheng Huang (Y.Huang-51@student.tudelft.nl), Xingyu Han (X.Han-5@student.tudelft.nl), Anyan Huang (A.Huang-3@student.tudelft.nl)
Affiliations:
Delft University of Technology

RainMamba

Title: RainMamba: A Video Coarse-to-Fine Mamba for Video Raindrop Removal
Members: Hongtao Wu[1] (wuhongtao@westlake.edu.cn), Hong Wang[2], Yefeng Zheng[1]
Affiliations:
[1]Medical Artificial Intelligence Laboratory, West Lake University
[2]School of Life Science and Technology, Xi'an Jiaotong University

RainDropX

Title: Leveraging Perceptual and Structural Constraints for Dual-Focused Raindrop Removal Using Restormer
Members: Abhijeet Kumar (ee23d406@smail.iitm.ac.in), Aman Kumar, A.N. Rajagopalan
Affiliations:
Indian Institute of Technology Madras

Cidaut AI

Title: FracDeformer: Fractional Fourier Aware Deformable Transformer for Day and Night Raindrop Removal
Members: Marcos V. Conde (marcos.conde@uni-wuerzburg.de), Paula Garrido, Daniel Feijoo, Juan C. Benito
Affiliations: Cidaut AI

DGL_DeRainDrop

Title: GSA2Step
Members: Guanglu Dong (dongguanglu@stu.scu.edu.cn), Xin Lin, Siyuan Liu, Tianheng Zheng, Jiayu Zhong, Shouyi Wang, Xiangtai Li, Lanqing Guo, Lu Qi and Chao Ren
Affiliations:
Sichuan University

xdu_720

Title: Raindrop Mask Prior-Guided Deraining Transformer
Members: Shuaibo Wang (shbwang@stu.xidian.edu.c), Shilong Zhang, Wanyu Zhou, Yunze Wu, Qinzhong Tan
Affiliations: Xi'dian University

EdgeClear-DNSST Team

Title: EdgeClear-DNSST: Edge-preserving Day-Night Sparse-Sampling Transformer
Members: Jieyuan Pei (peijieyuan@zjut.edu.cn), Zhuoxuan Li
Affiliations:
Zhejiang University of Technology & Tongji University

MPLNet

Title: MPLNet: Multi-Stage Progressive Learning with Illumination-Conscious Dynamic Transformer Networks
Members: Jiayu Wang (2024090902024@std.uestc.edu.cn)), Haoyu Bian, Haoran Sun
Affiliations:
University of Electronic Science and Technology of China

Singularity

Title: FracDeformer: Fractional Fourier Aware Deformable Transformer for Day and Night Raindrop Removal
Members: Subhajit Paul (subhajitpaul@sac.isro.gov.in)
Affiliations:
Space Applications Centre (SAC)

VIPLAB

Title: Staged restoration: deblurring first, then detail enhancement
Members: Ni Tang (23020240157683@stu.xmu.edu.cn), Junhao Huang, Zihan Cheng, Hongyun Zhu, Yuehan Wu
Affiliations:
School of Informatics, Xiamen University

2077Agent

Title: DiT-RainRemoval: Fine-Tuning a Pre-Trained DiT Model for Dual-Focus Raindrop Removal
Members: Kaixin Deng[1] (kaixin.deng.t0@elms.hokudai.ac.jp), Hang Ouyang[2], Tianxin Xiao[2], Fan Yang[2], Zhizun Luo[2]

Affiliations:
[1]Hokkaido University
[2]Chengdu University of Technology

X-L

Title: Permuted Self-Attention Network for Image Raindrop Removal
Members: Zeyu Xiao[1] (zeyuxiao1997@163.com), Zhuoyuan Li[2]
Affiliations:
[1] National University of Singapore
[2] University of Science and Technology of China

UIT-SHANKS

Title: Effective Raindrop Removal: An Integrated Deep Ensemble Approach with Semi-Supervised Learning and Restormer Denoising
Members: Pham Hoang Le Nguyen[1] (22520982@gm.uit.edu.vn), Dinh Thien An, Luu Thanh Son, Kiet Van Nguyen
Affiliations:
University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam

One Go Go

Title: Deraining and defocus-blurring separately
Members: Ronghua Xu [1] (202424131014T@stu.cqu.edu.cn), Xianmin Tian, Weijian Zhou, Jiacheng Zhang
Affiliations:
School of Big Data and Software Engineering, Chongqing University

DualBranchDerainNet

Title: A Lightweight Dual-Branch Raindrop Removal Method for Day and Night Scenes
Members: Yuqian Chen (213240247@seu.edu.cn)
Affiliations:
Southeast University

QWE

Title: Optimization of raindrop Image Restoration Model Integrating Data Enhancement and Continuous Learning
Members: Yihang Duan (2432179908@qq.com), Yujie Wu
Affiliations:
Xidian University

Visual and Signal Information Processing Team

Title: UFormer-based Dual-Focus Raindrop Removal for NTIRE 2025
Members: Suresh Raikwar (suresh.raikwar@thapar.edu), Arsh Garg, Kritika
Affiliations:
None

The Zheng family group

Title: No title
Members: Jianhua Zheng (460919144@qq.com), Xiaoshan Ma, Ruolin Zhao, Yongyu Yang, Yongsheng Liang, Guiming Huang
Affiliations:
Zhongkai University of Agriculture and Engineering

RainClear Pioneers

Title: RainClearNet: A Dual-Stream Physics-Guided Framework for Raindrop Removal
Members: Qiang Li (li_qiang@shu.edu.cn), Hongbin Zhang, Xiangyu Zheng
Affiliations:
Shanghai University

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 14

[2] Juan C Benito, Daniel Feijoo, Alvaro Garcia, and Marcos V Conde. Flol: Fast baselines for real-world low-light enhancement. *arXiv preprint arXiv:2501.09718*, 2025. 11

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Global context networks. *TPAMI*, 45(6):6881–6895, 2020. 12

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 4

[5] Wenhui Chang, Hongming Chen, Xin He, Xiang Chen, and Liangduo Shen. Uav-rain1k: A benchmark for raindrop removal from uav aerial imagery. In *CVPR*, pages 15–22, 2024. 2

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 15

[7] Hongming Chen, Xiang Chen, Jiyang Lu, and Yufeng Li. Rethinking multi-scale representations in deep deraining transformer. In *AAAI*, pages 1046–1053, 2024. 2

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 13

[9] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33. Springer, 2022. 7, 11, 13

[10] Xiang Chen, Jinshan Pan, Jiangxin Dong, and Jinhui Tang. Towards unified deep image deraining: A survey and a new benchmark. *arXiv preprint arXiv:2310.03535*, 2023. 2

[11] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 22367–22377, 2023. 5

[12] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 4

[13] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25627–25636, 2024. 2, 5

[14] Zheng Chen, Zongwei Wu, Eduard Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, Hongyuan Yu, Cheng Wan, Yuxin Hong, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In *CVPR*, pages 6108–6132, 2024. 4

[15] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *European Conference on Computer Vision*, pages 53–71. Springer, 2022. 5

[16] Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. *arXiv preprint arXiv:2412.13443*, 2024. 11

[17] Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N. Angelopoulos, and Ion Stoica. Prompt-to-leaderboard, 2025. 10

[18] Ning Gao, Xingyu Jiang, Xiuhui Zhang, and Yue Deng. Efficient frequency-domain image deraining with contrastive regularization. In *European Conference on Computer Vision*, pages 240–257. Springer, 2024. 7

[19] Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12097–12107, 2023. 14

[20] Yeying Jin, Xin Li, Jiadong Wang, Yan Zhang, and Malu Zhang. Raindrop clarity: A dual-focused dataset for day and night raindrop removal. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2, 11, 12

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 8, 11

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

[23] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 12, 13

[24] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *CVPR*, pages 5886–5895, 2023. 9

[25] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 8

[26] Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Promptcir: blind compressed image restoration with prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6442–6452, 2024. 5, 6

[27] Xin Li, Yeying Jin, Xin Jin, Zongwei Wu, Bingchen Li, Yufei Wang, Wenhan Yang, Yu Li, Zhibo Chen, Bihan Wen, Robby Tan, Radu Timofte, et al. NTIRE 2025 challenge on day and night raindrop removal for dual-focused images: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 6

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 4

[29] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *CVPR*, pages 2773–2783, 2024. 4

[30] Pengju Liu, Hongzhi Zhang, Jinghui Wang, Yuzhi Wang, Dongwei Ren, and Wangmeng Zuo. Robust deep ensemble method for real-world image denoising. *Neurocomputing*, 512:1–14, 2022. 9

[31] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 14

[32] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10346–10357, 2023. 9

[33] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 11

[34] Subhajit Paul, Sahil Kumawat, Ashutosh Gupta, and Deepak Mishra. F2former: When fractional fourier meets deep wiener deconvolution and selective frequency transformer for image deblurring. *arXiv preprint arXiv:2409.02056*, 2024. 13

[35] Subhajit Paul, Sahil Kumawat, Ashutosh Gupta, and Deepak Mishra. F2former: When fractional fourier meets deep wiener deconvolution and selective frequency transformer for image deblurring. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 13

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 9, 13, 14

[37] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 11

[39] Qiyu Rong, Hongyuan Jing, Mengmeng Zhang, Jinlong Li, and Mengfei Han. STRRNet: Semantics-guided two-stage raindrop removal network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 14

[41] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 7

[42] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *European Conference on Computer Vision*, pages 111–129. Springer, 2024. 8

[43] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *ECCV*, pages 111–129. Springer, 2024. 3

[44] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2353–2363, 2022. 15

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 11

[46] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 12

[47] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *ACMMM*, pages 7881–7890, 2024. 10

[48] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022. 8

[49] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 12

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 3, 5, 6, 7, 8, 11, 12, 15

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 11

[52] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 14

[53] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 12

[54] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, pages 12780–12791, 2023. 14

[55] Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. *arXiv preprint arXiv:2503.10622*, 2025. 12